人文科学とコンピュータ　41－2
（１９９９．１．２２）

新聞記事の文章末表現における特異的名詞語彙の出現特性

本間　愛　　中村　隆志
a-honma@dolphin.ge.niigata-u.ac.jp　　tks@human.ge.niigata-u.ac.jp
新潟大学人文学部

新聞記事の文章末に着目し、文書の文字数が大きくなるほど、文章末に出現する頻度が高くなるような名詞が存在するかどうか調査する。資料として、新聞記事データベース１年分から、キーワード検索を用いて得られる文書群の内、文書数が 500 から 1500 までの文書群を全て取り上げる。文書群の数は 495 個である。ある文書群に於いて、文書の文字数が増加するに連れ、文章末での出現頻度が増加する名詞を、その文書群の"大河ターム"という造語で呼称する。495 個の文書群からの抽出の結果、480 個(～ 97.0%)の文書群からのべ 4158 個の大河タームが取り出される。大河タームの 20%弱は、抽出に用いた文書群の頻出名詞であることがか確認される。複数の文書群から重複して抽出される回数が大きい大河タームの内、いくつかは、他の文書群での頻出名詞になることがなかったこと、つまり、長い文章の締め括り部分で使用され易い名詞が存在することが確認される。

キーワード：新聞記事、テキスト分類、文章末表現

Frequency of particular noun usage in closing sentences in newspaper articles

Ai Honma　　Takashi Nakamura
a-honma@dolphin.ge.niigata-u.ac.jp　　tks@human.ge.niigata-u.ac.jp
Department of Humanities, Niigata University

We focus on the frequency of noun usage in the closing sentences of a group of newspaper articles. We investigate the existence of terms whose usage frequency in the closing sentences increases as the article length increases. Data are compiled from 495 groups of newspaper articles obtained from the Nikkei Newspaper Database '97 which include more than 300 and less than 1500 articles. The nouns are called "taiga terms" whose usage frequency in the closing sentences of each article increases as the article length increases ("taiga" means 'big river' in Japanese). 4158 "taiga terms" were extracted from 480 article groups. A little less than 20% of "taiga terms" are the most frequently used nouns. Some "taiga terms" are repeatedly extracted from many groups of newspaper articles, and they are not the most frequent nouns in any groups of newspaper articles.

keywords: newspaper articles, text categorization, closing sentences

１：序に変えて

本稿の狙いは新聞記事の文書において、結びに使われ易い名詞を抽出することにある。新聞記事は自然言語処理研究の資料として、言語処理学会、情報処理学会、人工知能学会などの論文や報告書に頻繁に調査されている。自然言語処理において、文書の結びの部分、すなわち文章末表現は、大きく分けて２通りの扱いが為されている。

１つは新聞記事の主題推定に関する研究である[1]。この文献では、主題推定のためにテキスト構造を利用する可能性に触れているが、新聞記事の文章末では主題と関連性が小さいと推察されている。見出し語の大半が、記事本文の冒頭から数十単語の間に出現することが確かめられ、この結果が主題と新聞記事の文章末との関連性の希薄さを推察する根拠となっている。

もう１つは重要文抽出に関する研究である[2]。この文献で述べられる重要文自動抽出法は、抽出派でヒューリスティック型と位置づけされる。この方法は、要約文選択アルゴリズムの開始ステップにおいて、文章の最終文（及び冒頭文）を要約文として採用する。その根拠として、文章の最終文は著者が話を締め括る必要があるため、文章の他の部分と異なった、何らかの特別な意図を持って書かれた文であり、要約作成時にも欠かせない重要な文であるため、としている。

主題と関連性が希薄なことと、文章全体において先験的に重要な文であると位置づけすることは、必ずしも矛盾することではない。しかし、最終文あるいは文章末表現については、自然言語処理における一定の見解はないように思われる。この状況を踏まえて、中村他[3]は新聞記事のいくつかのテーマの文書群において、締め括りに使われ易い名詞を抽出する方法を提案し、その方法で得られる名詞を「大河ターム」と呼称した。この文献では、大河タームの抽出方法の提案と予備調査の報告に留まった。本稿では、この方法を日本経済新聞CD-ROM97 年度版[4]を網羅的に調査した結果を報告する。

２：文章作成指南書と自然言語処理

新聞記事の多くは、文章作成の訓練をある程度受けた者の手による。調査すべき新聞記事の文章をもたらす要因の１つとして、文章作成を教授する書物を改めて考察し、前章の自然言語処理研究のスタンスとの対応を述べる。

文章作成の作法を指南する書物は、受験参考書の小論文対策参考書や書店に並ぶ多くの出版社の新書、ハードカバーに至るまで、枚挙にいとまがない程である。これらに共通する主張は「他人が読む価値があるものをつくる」という心構えである。この心構えを実現するための作法は大きく分けて２つに分類され、その違いは作成すべき文章の種類や狙いの違いに反映される。

１つ目に分類されるのは、いわゆる実用的文章、すなわちレポート、報告書、論文の書き方を指南するものである。ここで「実用的」というタームはレポート、報告書、論文などを総括した文書分類のための用語であり、それらの文章作成指南書の中で使用されているものである。代表的なものとして篠田義明「コミュニケーション技術」[5]を例にとる。ここでは、ワンワード／ワンミーニング、ワンセンテンス／ワンアイディア、ワンパラグラフ／ワントピックの思想が徹底されており、細部に至るまで明晰さを重んじる。パラグラフ内では総論から各論へという作法が述べられ、パラグラフ内の文の順序は総論を先頭とし、各論

については、読み手の関心をより多く引くものから順に並べていくことを説く。２つの例外（総論の繰り返しによる説得法とパラグラフ内で因果関係を完結的に記述する方法）を除き、パラグラフの締め括り方への言及はない。また、文章全体の構成つまりパラグラフの連結については、序論、本論、結論の構成を推奨する（文献[6]においては、起承転結の構成を使用することについて、詩文作成の法則としての有用性を認めながらも、実用的文章に応用することを禁じている）。結論を述べるべき最終パラグラフも、一つのパラグラフに過ぎない。もしも、パラグラフ内における順序立て、すなわち、総論から各論へ、各論は関心を引くもの順に、という指南に乗っ取るならば、文章の締め括りは、結論パラグラフ内の、相対的に最も関心が低い各論の一つが述べられることになる。

もう一つは、実用的でない文章を含む一般的文章、エッセイや評論の書き方の指南書の類である。文献[7,8]では、文章全体の構成として「起承転結」を推奨する。読み手の関心を引くための「転」の展開の善し悪しを論じているが、「転」のためだけの議論ではない。「転」を持ち込むことは、文章構成に流れと立体感を与えるためであり、「結」の部分への接続を用いて読者に意外性を与え、面白がらせるためである。文章の締め括り方に具体的指南はないものの、読み手の関心を最後まで引き続け、締め括りの部分で面白がらせることを良しとする考えがある。また、文献[9]では、「均衡のとれた文章」「文章の遊び」「文章の品格」などのテーマに具体例に触れて解説しており、文献[10]では、「文章のリズム」の必要性を実例を挙げながら説いている。いずれも具体的作成方法についての指南はないものの、読者を飽きさせず、最後まで関心を引き続けることを良しとする考えに基づいている。文章の結び方に関しては、読み手への印象を考慮せよ、との心構えだけを説くものがほとんどであるが、締め括り部分が文章、あるいは書き手の評価に大いに影響があることを強調している点が、実用的でない文章の作成指南書の特徴である。

１章で述べた、自然言語処理での文章末の扱いの違いと、本章で述べた文章作成指南書の分類との対比をつけることが可能である。実用的文章作成指南に乗っ取り、総論から各論へ、各論の登場順位は関心の高いものから、という方針を守った文章であるならば、主題となる見出し語が文章の最初の方に出現することは明らかである。実用的でない文章指南に乗っ取り、読者の関心を最後まで引き続け、締め括り部分で読者に強い印象を与えようとした文章であるならば、最終文をアプリオリに重要文と見なすことは必然的である。作成すべき文章の性質とその作成指南法の違いが、自然言語処理での文章末に対する見解の違いとなって現れると考えられる。この文章の性質と作成指南法の違いを利用して、文章末に特徴的な表現（ここでは、特に名詞に限定する）を抽出するのが、本稿の狙いである。抽出のための仮説を次章で、方法を次々章以降で述べる。

３．仮説と方針

本稿では、新聞記事データベースから得られた文書群を資料として、文章末に用いられ易い表現、特に名詞の抽出を試みる。抽出のために一つの仮説を立てる。キーワード検索によって得られた文書群をマクロ的に捉えた場合、文字数の小さい文章ほど、実用的文章作成指南に乗っ取って書かれたものが多く、逆に文字数の大きい文章ほど、実用的文章の構成から外れ、実用的でない

文章の性質を帯び易くなる。つまり、文字数が大きい文章になればなる程、締め括り部分において、その文章のテーマの上位概念、最も主張すべき意見を代表するターム、そのテーマを取り巻く状況の今後の指針、などを述べる割合が多くなるだろうという仮説である。

上述の仮説の根拠を述べる。文字数が200 字程度、あるいはセンテンス数が４から５程度の文章では、起承転結の構成をとったり、リズムやバランスを考慮する必要が発生するほどの情報量があるとは考えにくく、この場合は実用的文章の構成をとることが多くなると考えられる。一方で、文字数が数千字程度、段落もいくつかとるような大きさの文章になると、空間的余地ができ、起承転結の構成をとり、リズムやバランスに配慮することができる。逆に数千字程度の長さになると、関心の高いもの低いものへと順に各論を述べて終わる様な文章では、最後まで読まれないだろう。よって、これらの文章では、最後まで読者の関心を引き続け、締め括りで読者に強い印象を与えるような文章が相対的に多くなると考えられる。個別の新聞記事を当たれば、上記の仮説に違うものもいくつかあることは充分考えられる。短くまとまっていて、締め括りが為されているものもあれば、社説やコラムの欄にある記事でも、明らかに締め括りに失敗しているものもあるだろう。しかし、マクロ的に見れば、空間的余地が実用的文章と実用的でない文章の割合と相関があると仮定して良いとみなす。

そこで、あるテーマに関連する文書群について、それらの締め括り部分、つまり最後の数センテンスを取り出した場合を考える。文字数の小さい文書においては、そのテーマのあるトピックについて各論が述べられていることが多いと考えられる。文字数の大きい文書においては、読者に強い印象を与えるような表現が多いと考えられる。これは文字数が大きくなるほど、はっきりした傾向として現れるだろう。つまり、文章末の数千テンスに注目し、そこでの使用頻度が文字数につれて高くなるタームがあるならば、そのテーマの文章の締め括りに使用され易いタームか、あるいは文章一般において使用され易いタームであることになる。これらのタームを取り出すための具体的方法を次章で述べる。

４．大河タームの抽出方法

以下の手順０から手順８を適用して抽出された名詞を"大河ターム"と呼ぶ[3]。

０．調査対象は新聞記事データベースを選び、本報告では日本経済新聞 CD-ROM、97 年版[4]を使用した。

１．キーワード検索を行い、検索結果の文書数が 500 個から 1500 個までの文書群を全て選び、各、得られた文書をまとめて、１つの文書群とする。

２．検索して得られた文書群のうち、センテンス数が４未満のもの、あるいは文章でなく箇条書きに終始するものを除外する。

３．声明で記事が終わる場合は、発言内外を問わず、句点から句点まで、あるいは鉤括弧までを１つのセンテンスとする。

４．文書群内の全ての文書を文字数順に並べ、ほぼ同数になるようにX個のブロックに分割する。平均文字数最小のブロックを第０ブロックとし、昇順に順序付けを行い、平均文字数最大のブロックを第$(X-1)$ブロックとする。

５．各文書群において、全ての文書の末尾のkセンテンスを取り出し、これらに対して、日本語形態素解析プログラム「茶筌 1.0」を用いて形態素解析を行う[11]。こ

の中から普通名詞、固有名詞、サ変名詞、時相名詞を取り出し、各名詞毎に集計を行う。

6．ある文書群 j において、第 x ブロック内の全ての文書の末尾 k センテンスにおける、ある名詞 n の出現回数を $F(j,n,x)$ とする。$F(j,n,x)$ の最大値を $Fmax(j,n)$ とする。文書群 j 内の全文書数をＡ(j) とする。この時、$Fmax(j,n') \geqq A(j)/X/10$ を満たさない場合、名詞 n' を手順 7,8 で行う解析から除外する。

7．頻度分布 $(x,\ F(j,n,x))$ に対して、単純回帰分析を行い、回帰係数を求める。さらに、回帰係数の検定を行い、回帰係数の値が 0 である帰無仮説を有意水準 0.05 で棄却できる名詞のみを抽出する。また、得られた回帰係数 E が $E > t$ を満たすことを抽出の条件とする。

8．頻度分布 $(x,\ F(j,n,x))$ に対して、歪み度 $V(n)$ を求める。手順７で抽出された名詞の内、$V(n) < 0$ を満たすもののみを抽出する。

以上の手順により、文書群から抽出された名詞の集合をその文書群の大河タームと呼び、T-T と表す。各手順について解説する。

手順１：各の大河タームが末尾 k センテンスに出現する頻度を大きくするために、文書数が大きい文書群を選ばなければならない。500 から 1500 という数値はヒューリスティックな判断である。

手順２：センテンス数が４未満のものは、起承転結の構成を作りようがなく、締め括る必要があるほど多くの情報量に満ちていることは考えられない。また、今回の調査が末尾での文章の締め括りの調査であるため、箇条書きの文書は対象外とする。

手順３：鉤括弧を一つの特別なセンテンスとせず、鉤括弧内のセンテンスも、鉤括弧外のものと同様のセンテンスとする。

手順４：手順６以降の頻度分布を得るため、文書群を文字数順に並べて分割する。割り切れなくてもそのまま用いることにする。

手順５：解析対象として末尾の表現を取り出すにあたって、k センテンスにした理由は、１センテンスだとデータが少ないからであり、また、文章を締め括る機能が１センテンスだけに集約されていると限定されているよりは、k センテンスに拡がっていることが多いと推測するからである。名詞は具体的対象、抽象的表象が存在するもの、及び時間的表現をあらわすものに限定する。よって、形式名詞、数詞、副詞的名詞は除外する。なお、本章及び以下の章で述べる名詞とは、日本語形態素解析プログラム茶筌 1.0 の解析結果から求められた形態素を表す。

手順６：ここでは、最大値 $Fmax(j,n')$ の大きさを採否の条件とする。$A(j)/X/10$ を敷居値として採用するのはヒューリスティックな判断である。

手順７：単純回帰分析で得られた係数を意味づける指標として、回帰係数の値が０である帰無仮説を棄却することの他に、決定係数の値があるが、ここでは採用しない。頻度分布の変化は、例えば級数的変化でも良く、必ずしも、線形回帰直線で近似される必要はない。増加傾向にあるか、減少傾向にあるかの意味づけを必要とするため、回帰係数がゼロである帰無仮説を棄却するものだけを抽出する。ただし、回帰係数の値が小さい場合は、抽出から除外する。

手順８：頻度分布 $(x,\ F(j,n,x))$ の増加傾向を裏付けするための指標として、歪み度を用いる。歪み度が負の値をとる時、頻度分布が横軸正方向に偏っていることを示す性質を利用している。手順７での抽出を裏付けするための指標である。

抽出された大河タームとの比較のために各の文書群から２つのテーブルを抽出する。各の文書群全体の最頻出名詞の上位 w 個を取り出して得られるテーブルを OTF、各の文書群の文書の末尾 k 文のみから名詞の頻度を求め、その最頻出名詞の上位 w 個を取り出して得られるテーブルを OPF と呼ぶ。

５．抽出結果

４章で述べた手順の中のパラメータは、X=5, k=2, t=1.0 を用いた。

１：上記の方法の手順１により、495 個の文書群が得られた。この 495 個の文書群に対して、手順２から手順８までを実行した結果、480 個(～ 97.0%)の文書群から大河タームが得られた。大河タームが抽出されなかった 15 個の文書群の検索のためのキーワードはそれぞれ、「イベント」、「飲食業」、「家族」、「観光客」、「観光地」、「個人消費」、「国際紛争」、「地球」、「展示会」、「表彰」、「部屋」、「福島」、「文化施設」、「来日」、「惑星」、であった。

２：480 個の文書群から得られた大河タームはのべ 4158 個であった。文書群毎の大河タームの個数の平均値は～ 8.45、メジアンは 7、標準偏差は～ 5.86 であった。

３：文書群毎の大河タームの個数の平均値とメジアンの値から、w=8 として OTF、OPF を作成し、大河タームと比較する。各の大河タームについて、その文書群の OTF に含まれるものは 4158 個中 747 個(～ 18.0%)、その文書群の OPF に含まれるものは 4158 個中 732 個(17.6%)、OTF、OPF の両方に含まれるものは 4158 個中 596 個(14.3%)であった。

４：同じ大河タームが複数の文書群から重複して抽出されることが起こる。ある名詞の重複抽出の回数が n 回なら、その名詞は n 個の文書群で大河タームとして抽出されたことになる。480 個の文書群から抽出された全ての大河タームについて、重複抽出の回数をカウントした。最大個数のもの

大河ターム	重複回数	大河ターム	重複回数
企業	168	競争	49
日本	161	米	49
今後	132	可能性	44
市場	121	銀行	36
問題	67	金融	35
経済	66	アジア	34
改革	57	開発	34
経営	57	技術	31
会社	55	情報	31
事業	54	地域	30

表１．重複抽出回数の大きい大河ターム

大河ターム	重複回数	大河ターム	重複回数
今後	132	検討	12
可能性	44	声	12
業界	24	自治体	11
政治	19	金融機関	10
見方	17	今回	10
考え	16	手	10
コスト	15	各社	9
大手	14	関係	8
国内	13	程度	8
社会	13	影響	7

表２．他の文書群の OTF に含まれない大河タームで重複抽出回数の大きいもの

から上位20個を表1に示す。

5：文書群毎の大河タームは他のキーワード検索で得られる文書群においては頻出名詞となって OTF に含まれることは大いにあり得る。逆に全ての文書群の OTF に含まれないが、大河タームとして複数の文書群から重複して抽出されるようなものも存在する。全ての文書群の OTF に含まれず、かつ重複抽出の回数の多い大河タームの上位20個を表2に示す。

6．結論

結果1から日本経済新聞 CD-ROM97 年版の記事数が 500 から 1500 である文書群の内、約 97%の文書群から大河ターム、すなわち文書の文字数が大きくなるに連れて結びの2文に出現する頻度が高くなる名詞が存在することが確認された。結果2から抽出された大河タームは文書群毎に平均すれば約 8.45 個であったが、ばらつきが大きい。結果3から上位8位までの OTF、OPF に大河タームが含まれる割合を求めたが、20%弱のものが該当した。これらは、文字数が大きくなるに連れて結びの2文に出現する頻度が高くなるような名詞でありながら、その文書群全体の最頻名詞でもあること、あるいは文書群内の全ての文書の結び2文だけに注目した頻度調査においても上位に現れる様な頻出名詞である。文字数の大きい文書の結び2文に於いては、OTF、OPF などに含まれるような最頻出名詞と、頻出しない名詞の両方が使われ易いことを示している。

結果4から得られる表1と結果5から得られる表2とを比べると、表1と表2の両方に、「今後」「可能性」が含まれている。表2に現れる大河タームは、今回の資料の全ての文書群において OTF として最頻名詞にならなかったものである。逆に表1に現れながら、表2に含まれないものは他の文書群のいずれかで頻出する名詞であることになる。同じ事象を語るにしても、様々な視点があるならば、ある文書群の大河タームが別の文書群では何度も問題になるような物象を指す名詞であったり、そのテーマを構成するような中心的概念であることは大いにあり得ることである。一方で、ある文書群の大河タームが全ての文書群で頻出名詞にあたらないものを考えよう。それらは、文字数が大きい文書の結びに使われ易い名詞でありながら、特定のテーマの構成を担うことがほとんどない名詞である、つまり文字数が大きい文書の結び以外の部分では頻繁には使われず、結びの部分で多く使われ易い名詞であることになる。日本経済新聞 CD-ROM97 年版内の新聞記事全体を通して、「今後」「可能性」を含む表2に掲げた名詞が、その最たるものとなる。

7．課題

今後の課題を展望しておく

1：最頻出名詞と大河タームの意味上の関係を解析する。

2：大河タームのない文書群の文書の特徴を、大河タームのあるものと比較して解析する。

3：同じ文書群の中でも、抽出された大河タームを用いて締め括る文書と使わずに締め括る文書がある。それらの読後感の違いを被験者を用いた評価実験を行う。

謝辞

日本語形態素解析システム茶筌を開発し、精度向上に取り組み続けている奈良先端科学技術大学院大学松本研究室の皆様に謝意を申し上げます。また、調査を許可下さった日本経済新聞社殿に感謝いたします。

参考文献

[1] 野本忠司、松本裕治：テキスト構造を利用した主題の推定について、情報処理学会報告、NL114-8,1996.

[2] 山本和英、増山繁、内藤昭三：文章内構造を複合的に利用した論説文要約システム GREEN、自然言語処理, vol. 2(1), pp. 39-55,1995.

[3] 中村隆志、本間愛：新聞記事における文章末表現での名詞語彙の出現特性、情報処理学会報告、NL128-13,1998

[4] 日本経済新聞社、日本経済新聞 97 年 CD-ROM 版、日本経済新聞社、1997.

[5] 篠田義明：コミュニケーション技術、中公新書 807、1986.

[6] 澤田昭夫：論文の書き方、講談社学術文庫、1977

[7] 向井敏：文章読本、文藝春秋、1988.

[8] 中西一弘編：基礎文章表現法、朝倉書店、1996

[9] 辰濃和男：文章の書き方、岩波新書、1994

[10] 本田勝一：日本語の作文技術、朝日文庫、1982

[11] 松本裕治、北内啓、山下達雄、平野喜隆、今一修、今村友明：日本語形態素解析システム『茶筌』version 1.0 使用説明書、Information Science Technical Report, NAIST-IS-TR97007, 奈良先端科学技術大学、1997.